J Math Chem (2010) 48:1010–1035
DOI 10.1007/s10910-010-9722-2

ORIGINAL PAPER

D-MORPH regression: application to modeling with unknown parameters more than observation data

Genyuan Li · Herschel Rabitz

Received: 2 June 2010 / Accepted: 18 August 2010 / Published online: 2 September 2010

Abstract Diffeomorphic modulation under observable response preserving homotopy (D-MORPH) is a model exploration method, originally developed for differential equations. We extend D-MORPH to regression treatment of a model described as a linear superposition of basis functions with unknown parameters being the expansion coefficients. The goal of D-MORPH regression is to improve prediction accuracy without sacrificing fitting accuracy. When there are more unknown parameters than observation data, the corresponding linear algebraic equation system is generally consistent, and has an infinite number of solutions exactly fitting the data. In this case, the solutions given by standard regression techniques can significantly deviate from the true system structure, and consequently provide large prediction errors for the model. D-MORPH regression is a practical systematic means to search over system structure within the infinite number of possible solutions while preserving fitting accuracy. An explicit expression is provided by D-MORPH regression relating the data to the expansion coefficients in the linear model. The expansion coefficients obtained by D-MORPH regression are particular linear combinations of those obtained by least-squares regression. The resultant prediction accuracy provided by D-MORPH regression is shown to be significantly improved in several model illustrations.

Keywords D-MORPH · Least-squares regression · Regularization · Ridge regression · Smoothing splines · Orthonormal polynomial

G. Li · H. Rabitz (✉)
Department of Chemistry, Princeton University, Princeton, NJ 08544, USA
e-mail: hrabitz@princeton.edu

G. Li
e-mail: genyuan@princeton.edu

1 Introduction

In many modeling applications, the number of sought after parameters is larger than the amount of observed data. For example, in radiology and biomedical imaging, far fewer measurements are performed than the number of pixels in the desired image [1–3]. In genomics, a modest number of observations may be available compared to the total number of genes involved [4]. This situation also occurs in industrial and engineering applications. For instance, the ignition time of a fuel depends on pressure, temperature and the equivalence ratio for the fuel and oxygen [5], and in many circumstances, only a limited number of measurements are readily available. Determination of a model in such settings is a challenge, and progress in this regard could have a significant impact in many areas of science and engineering [6–8].

A mathematical model generally is an approximation to the structure of a complex physical system. Imperfect knowledge of the physical system, restrictions to the mathematical description of complex systems, and limited amounts of available data can make the resultant model predictions of unacceptable quality. The relationship between model prediction error, model complexity and data set size is a subject that continues to receive much attention [9,10].

Consider a system described by

$$y = f(\mathbf{x}) + \varepsilon, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ are the inputs, y is the output with random error ε (i.e., $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$). Here, the inputs $\mathbf{x}$ are assumed to be exactly measured. The expected prediction error of a regression fit (i.e., a model) $\hat{f}(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_0$ for squared-error can be represented as

$$\begin{aligned} \mathrm{Err}(\mathbf{x}_0) &= \sigma_\varepsilon^2 + [E(\hat{f}(\mathbf{x}_0)) - f(\mathbf{x}_0)]^2 + E[\hat{f}(\mathbf{x}_0) - E(\hat{f}(\mathbf{x}_0))]^2 \\ &= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(\mathbf{x}_0)) + \mathrm{Var}(\hat{f}(\mathbf{x}_0)). \end{aligned} \tag{2}$$

The first term, irreducible error, is the variance of the output y around its true value $f(\mathbf{x}_0)$ and cannot be avoided regardless of how well we estimate $f(\mathbf{x}_0)$, unless $\sigma_\varepsilon^2 = 0$. The second term, the bias of the model, is the square of the difference between the average value of the model obtained from different sample sets and the true value of the system. The bias measures how far away the model is from the true system. The last term represents the variance of the model determined from different sets of data. Two circumstances will be considered in this paper: (i) y is an experimental datum in which $\sigma_\varepsilon^2 \neq 0$ and (ii) y is given from numerical simulation of a complex system and $\sigma_\varepsilon^2 = 0$. In both cases, the goal is to prescribe a model $\hat{f}(\mathbf{x})$, generally nonlinear in $\mathbf{x}$, that faithfully represents the input→output ($\mathbf{x} \rightarrow y$) behavior of the original system.

Increasing the model complexity (e.g., by including more terms in the mathematical model $\hat{f}(\mathbf{x})$) generally will reduce model bias and offer a better fit to training data, but may increase the variance of the model (i.e., improving the accuracy of fitting the training data set by increasing model complexity can cause a reduction in prediction

accuracy). To have good prediction accuracy, traditionally a trade-off between model bias and variance is made through the use of models with modest complexity.

As a common special case of Eq. 1, consider a model with output y composed of a linear combination of fixed basis functions $\phi_i(\mathbf{x})$

$$y(\mathbf{x}, \mathbf{w}) = \hat{f}(\mathbf{x}) = \sum_{i=1}^{m} w_i \phi_i(\mathbf{x}), \tag{3}$$

here $\phi_1(\mathbf{x}) = 1$, and the expansion coefficients $\mathbf{w} = (w_1, w_2, \ldots, w_m)^T$ are constant parameters to be identified. The parameters $\mathbf{w}$ are often determined from a set of observation data $(\mathbf{x}^{(j)}, y^{(j)})$ $(j = 1, 2, \ldots, N)$ from the equations

$$\phi(\mathbf{x}^{(j)})^T \mathbf{w} = y^{(j)}, \quad (j = 1, 2, \ldots, N) \tag{4}$$

or in matrix form

$$\mathbf{\Phi w} = \mathbf{y}, \tag{5}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}^{(1)}) & \phi_2(\mathbf{x}^{(1)}) & \ldots & \phi_m(\mathbf{x}^{(1)}) \\ \phi_1(\mathbf{x}^{(2)}) & \phi_2(\mathbf{x}^{(2)}) & \ldots & \phi_m(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}^{(N)}) & \phi_2(\mathbf{x}^{(N)}) & \ldots & \phi_m(\mathbf{x}^{(N)}) \end{bmatrix} \tag{6}$$

and

$$\mathbf{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})^T. \tag{7}$$

Importantly, although the unknown coefficients $\mathbf{w}$ enter linearly in the model of Eq. 3, the relationship $\mathbf{x} \rightarrow y$ can be nonlinear through the basis functions $\phi_i(\mathbf{x})$.

Least-squares regression is commonly used to determine $\mathbf{w}$ by minimizing the residual sum of squares

$$\mathrm{RSS} = \|\mathbf{\Phi w} - \mathbf{y}\|^2. \tag{8}$$

In order to reduce the model variance, a variety of methods have been proposed including ridge regression [11–16], smoothing splines [17–19], etc. In ridge regression a regularization term is included in the minimization:

$$\mathrm{RSS} = \|\mathbf{\Phi w} - \mathbf{y}\|^2 + \|\Gamma \mathbf{w}\|^2 \tag{9}$$

for some suitably chosen Tikhonov matrix, Γ. In many cases, Γ is chosen as proportional to the identity matrix $\Gamma = \lambda^{1/2}\mathrm{I}$ with parameter $\lambda(>0)$, giving preference to solutions $\mathbf{w}$ with smaller norms:

$$\mathrm{RSS} = \|\mathbf{\Phi w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2. \tag{10}$$

Smoothing splines introduce a second term to penalize the square curvature norm of the model function. These methods generally pose a trade-off between the two terms with the following shortcomings: (1) in order to improve prediction accuracy, then fitting accuracy often has to be sacrificed to some degree, (2) determination of the optimal value for the parameter λ (e.g., by the discrepancy principle, cross-validation, the L-curve method, restricted maximum likelihood and unbiased predictive risk estimator validation) [20] often requires extensive computational effort. This paper introduces the D-MORPH (Diffeomorphic Modulation under Observable Response Preserving Homotopy) regression technique for modeling with more unknown parameters than observation data. The goals of D-MORPH regression are to: (1) reduce model variance while preserving fitting accuracy, and (2) avoid determination of regularization parameters, like λ in Eq. 10.

D-MORPH is a general approach for model exploration, which was originally established for treating quantum control applications [21–23], and has been extended to performing a homotopy-based directed exploration of models in other circumstances [24]. In modeling, a suitable mathematical form is often known, and the goal is to select the best model within the form. It is also possible that better models exist beyond the known form. D-MORPH allows for systematic incorporation of a wide variety of information about the system global features while searching for the best consistent model that preserves desired features and diminishes undesired properties.

Here, we apply D-MORPH to linear basis function regression. Although the form of the basis functions $\phi_i(\mathbf{x})$'s is fixed, increasing the number of basis functions can be included as desired. When the number of unknown parameters $\mathbf{w}$ is larger than the number of observation data in Eq. 5, there are an infinite number of solutions $\mathbf{w}$ exactly fitting $\mathbf{y}$. All the solutions $\mathbf{w}$ comprise a linear manifold $\mathcal{M}$, and all the functions $\Phi\mathbf{w}$ with $\mathbf{w} \in \mathcal{M}$ are "consistent models". Then D-MORPH can be applied to search for the best solution of $\mathbf{w} \in \mathcal{M}$ minimizing certain undesired properties, analogous to regularization terms in ridge regression and smoothing splines, to improve prediction accuracy. When the data $\mathbf{y}$ are accurate (e.g., either from a reliable time-consuming numerical simulation, or believed to be accurately measured), we specify that D-MORPH regression exactly preserves $\mathbf{y}$. When the data $\mathbf{y}$ contain noise, exactly fitting $\mathbf{y}$ is not appropriate. In this case, the data can be filtered to produce a meta-data set $\hat{\mathbf{y}}$, which is then used as the preserved features in D-MORPH search for a consistent model. As will be demonstrated in this work, D-MORPH preserves fitting accuracy while improving prediction accuracy over existing methods.

The paper is organized as follows. In Sect. 2 the methodology of D-MORPH regression is presented with a simple illustration. In Sect. 3 the ignition data of an H_2/air combustion model with and without noise are treated by D-MORPH regression. Some concluding remarks are given in Sect. 4.

2 Methodology of D-MORPH Regression

2.1 General solution of a consistent linear algebraic equation system

Eq. 5 is a linear algebraic equation system for the variables in the vector $\mathbf{w}$. If the system is consistent, i.e.,

$$\mathbf{y} \in \text{Im } \mathbf{\Phi} = \{\mathbf{\Phi z} | \mathbf{z} \in \Re^m\} \tag{11}$$

where 'Im' denotes image, the solution for $\mathbf{w}$ exists, but it may not be unique.

The solution of Eq. 5 can be obtained by using the generalized inverse of $\mathbf{\Phi}$. For a $p \times q$ matrix A its generalized inverse is a $q \times p$ matrix G satisfying part or all of the Penrose conditions [25]:

$$\begin{array}{ll} (1)\ AGA = A, & (2)\ GAG = G, \\ (3)\ (AG)^T = AG, & (4)\ (GA)^T = GA. \end{array} \tag{12}$$

The generalized inverse A^+ for A satisfying all four Penrose conditions is unique. A $p \times q$ matrix A with rank $r (\leq \min(p, q))$ can be decomposed as

$$A = H \begin{bmatrix} R_r & 0 \\ 0 & 0 \end{bmatrix} K^T, \tag{13}$$

where H and K are $p \times p$ and $q \times q$ orthogonal matrices, respectively; and R_r is a nonsingular r-dimensional upper triangular matrix. A^+ is then represented as

$$A^+ = K \begin{bmatrix} R_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} H^T. \tag{14}$$

Various routines are available to calculate A^+ including within Matlab [26]. When A has a full row or column rank, A^+ has the following form:

$$A^+ = \begin{cases} A^T(AA^T)^{-1}, & \text{if } p < q \text{ and } A \text{ is of full row rank } p \\ (A^TA)^{-1}A^T, & \text{if } p > q \text{ and } A \text{ is of full column rank } q \\ A^{-1}, & \text{if } p = q \text{ and } A \text{ is of full row/column rank } p, q \end{cases} \tag{15}$$

Let $\mathbf{\Phi}^-$ be one of the generalized inverses of $\mathbf{\Phi}$ satisfying the first condition

$$\mathbf{\Phi \Phi^- \Phi} = \mathbf{\Phi}. \tag{16}$$

Then the general form for the m-dimensional vector $\mathbf{w}$ as the solution to Eq. 5 is

$$\mathbf{w} = \mathbf{\Phi^- y} + (\mathrm{I}_m - \mathbf{\Phi^- \Phi})\mathbf{v}, \tag{17}$$

where I_m is the identity matrix with dimension m and $\mathbf{v}$ is an arbitrary vector in $\Re^m$. One choice for $\mathbf{\Phi}^-$ in Eq. 17 is $\mathbf{\Phi}^+$, which is the solution $\mathbf{w}$ given by traditional least-squares regression, and the solution $\mathbf{w}$ obtained by D-MORPH regression is a particular linear combination of this solution.

2.2 Exploring a consistent linear basis function model satisfying additional requirements by D-MORPH regression

For the linear basis function regression model in Eq. 5, we seek a solution $\mathbf{w} \in \Re^m$. As $N < m$, there are an infinite number of solutions which comprise a submanifold $\mathcal{M} \subset \Re^m$. Since $\mathcal{M}$ is obtained from the linear restrictions in Eq. 5 on $\Re^m$, then $\mathcal{M}$ is a convex set and completely connected. We may consider Eq. 5 as a map $\mathcal{J}: \mathcal{M} \to \Re^N$ with each component $\mathcal{J}_j$

$$\mathcal{J}_j: \mathcal{M} \to \Re^1, \qquad 1 \le j \le N \tag{18}$$

where $\mathcal{J}_j = \phi(\mathbf{x}^{(j)})^T\mathbf{w}$. Within $\mathcal{M}$ the image of $\mathcal{J}_j$ is a constant $y^{(j)}$ which is a feature to be preserved in D-MORPH exploration.

Now consider an arbitrary exploration path $\mathbf{w}(s)$ within $\mathcal{M}$ with s in $[0, \infty)$. As the features $\mathbf{y}$ are preserved over the entire path, $\mathcal{J}$ is independent of s, i.e.,

$$\frac{\mathrm{d}\mathcal{J}}{\mathrm{d}s} = \mathbf{\Phi}\frac{\mathrm{d}\mathbf{w}(s)}{\mathrm{d}s} = \mathbf{0}, \tag{19}$$

which implies that

$$\frac{\mathrm{d}\mathbf{w}(s)}{\mathrm{d}s} \in \mathrm{Ker}\ \mathbf{\Phi} = \{\mathbf{z}|\mathbf{\Phi}\mathbf{z} = \mathbf{0}, \mathbf{z} \in \Re^m\}, \tag{20}$$

where 'Ker' denotes kernel. Eq. 19 is satisfied if $\mathrm{d}\mathbf{w}(s)/\mathrm{d}s$ is defined as

$$\frac{\mathrm{d}\mathbf{w}(s)}{\mathrm{d}s} = P\mathbf{v}(s) = (\mathrm{I}_m - G\mathbf{\Phi})\mathbf{v}(s) \tag{21}$$

because the image space of P is contained in Ker $\mathbf{\Phi}$, where G is a generalized inverse of $\mathbf{\Phi}$ satisfying Penrose condition (1), and $\mathbf{v}(s)$ is an arbitrary function vector in $\Re^m$.

It can be readily proved that P is a projector, i.e.,

$$P^2 = P. \tag{22}$$

If $\mathbf{\Phi}^+$ is used as G, then P is also symmetric,

$$P^T = P, \tag{23}$$

i.e., P is an orthogonal projector [25].

The free function vector $\mathbf{v}(s)$ not only enables broad choices for exploring $\mathbf{w}(s)$, but also provides the possibility of continuously reducing a defined cost (e.g., the model variance in Eq. 2, fitting smoothness, etc.) along the exploration path. Thus, define a map $\mathcal{K}: \mathcal{M} \to \Re^1$ as a cost function. Similar to $\mathcal{J}_i$ we have

$$\frac{\mathrm{d}\mathcal{K}(\mathbf{w}(s))}{\mathrm{d}s} = \left(\frac{\partial\mathcal{K}(\mathbf{w}(s))}{\partial\mathbf{w}}\right)^T\frac{\mathrm{d}\mathbf{w}(s)}{\mathrm{d}s} = \left(\frac{\partial\mathcal{K}(\mathbf{w}(s))}{\partial\mathbf{w}}\right)^T P\mathbf{v}(s). \tag{24}$$

If we choose the free function vector

$$\mathbf{v}(s) = -\frac{\partial \mathcal{K}(\mathbf{w}(s))}{\partial \mathbf{w}} \tag{25}$$

and use the properties of the orthogonal projector P in Eqs. 22 and 23, we have

$$\begin{aligned}\frac{\mathrm{d}\mathcal{K}(\mathbf{w}(s))}{\mathrm{d}s} &= -\left(\frac{\partial \mathcal{K}(\mathbf{w}(s))}{\partial \mathbf{w}}\right)^T P \frac{\partial \mathcal{K}(\mathbf{w}(s))}{\partial \mathbf{w}} \\ &= -\left(P\frac{\partial \mathcal{K}(\mathbf{w}(s))}{\partial \mathbf{w}}\right)^T \left(P\frac{\partial \mathcal{K}(\mathbf{w}(s))}{\partial \mathbf{w}}\right) \leq 0. \end{aligned} \tag{26}$$

The cost $\mathcal{K}$, used as an additional requirement, will be continuously reduced (i.e., systematically refining the model) over the course of exploring the model for $s \geq 0$. One choice for $\mathcal{K}$ is given in Sect. 2.3.

2.3 Cost function based on the weighted norm of $\mathbf{w}$ and second order derivatives of $y(\mathbf{x})$

The cost function is defined as

$$\mathcal{K} = \frac{1}{2}\mathbf{w}^T B \mathbf{w}, \tag{27}$$

where B is symmetric and nonnegative definite. For example, B may be a diagonal matrix with elements $B_{ii} = b_i \geq 0\,(i = 1, 2, \ldots, m)$. This means that $\mathcal{K}$ is a weighted norm of $\mathbf{w}$, i.e.,

$$\mathcal{K} = \frac{1}{2}\sum_{i=1}^{m} b_i w_i^2 \tag{28}$$

and the w_i coefficients will shrink under D-MORPH at different rates dictated by the b_i values. This cost function is especially useful to suppress over-fitting with polynomial basis functions $\phi_i(\mathbf{x})$'s. Larger values of b_i may be placed on higher degree polynomials to suppress their contributions and smooth the fitting.

As a special case of Eq. 27, the cost function $\mathcal{K}$ can be also defined in terms of the second order derivatives of $y(\mathbf{x})$ as

$$\begin{aligned}\mathcal{K} &= \frac{1}{2M}\sum_{p=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2 y(\mathbf{x}^{(p)})}{\partial x_i \partial x_j}\right)^2 \\ &= \frac{1}{2M}\sum_{p=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{w}^T D_{ij}\phi(\mathbf{x}^{(p)})(D_{ij}\phi(\mathbf{x}^{(p)}))^T\mathbf{w}\end{aligned}$$

$$= \frac{1}{2}\mathbf{w}^T \left[\frac{1}{M} \sum_{p=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}\phi(\mathbf{x}^{(p)})(D_{ij}\phi(\mathbf{x}^{(p)}))^T \right] \mathbf{w}$$
$$= \frac{1}{2}\mathbf{w}^T D\mathbf{w} \tag{29}$$

where $M(>0)$ is a large number to control the magnitudes of the elements of D, for instance, $M = mN$, and

$$(D_{ij}\phi(\mathbf{x}^{(p)}))^T = \left(\frac{\partial\phi_1(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \quad \frac{\partial\phi_2(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \quad \cdots \quad \frac{\partial\phi_m(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \right). \tag{30}$$

We may also add a different weight to each of the second order partial derivatives $\partial\phi_k(\mathbf{x}^{(p)})/\partial x_i \partial x_j$, for example

$$(D_{ij}\phi(\mathbf{x}^{(p)}))^T = \left(b_1^{1/2} \frac{\partial\phi_1(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \quad b_2^{1/2} \frac{\partial\phi_2(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \quad \cdots \quad b_m^{1/2} \frac{\partial\phi_m(\mathbf{x}^{(p)})}{\partial x_i \partial x_j} \right). \tag{31}$$

In the two cases of Eqs. 27 and 29, respectively, B and D are symmetric and nonnegative definite, and the free function vector is

$$\mathbf{v}(s) = -\frac{\partial\mathcal{K}}{\partial\mathbf{w}} = -C\mathbf{w}(s), \tag{32}$$

where C represents B or D. The differential equation for the path, Eq. 21, becomes

$$\frac{\mathrm{d}\mathbf{w}(s)}{\mathrm{d}s} = -(\mathrm{I}_m - \mathbf{\Phi}^+\mathbf{\Phi})C\mathbf{w}(s) = -PC\mathbf{w}(s). \tag{33}$$

The solution of Eq. 33 is

$$\mathbf{w}(s) = e^{-sPC}\mathbf{w}(0). \tag{34}$$

Using the Poincaré separation theorem for eigenvalues [27], it can be proved that when C is nonnegative definite and P is an orthogonal projector with rank $m - l$, the eigenvalues of PC satisfy the relation

$$\lambda_i(C) \geq \lambda_i(PC) \geq \lambda_{l+i}(C), \quad i = 1, 2, \ldots, m - l \tag{35}$$

where

$$\lambda_1(C) \geq \lambda_2(C) \geq \cdots \geq \lambda_{m-1}(C) \geq \lambda_m(C).$$

As $\lambda_i(C) \geq 0$, all the nonzero eigenvalues of PC are positive.

Generally, PC is not a projector. Suppose PC has rank r and can be decomposed as

$$PC = Q\Lambda Q^{-1} \tag{36}$$

where Q is the eigenvector matrix of PC which may not be an orthogonal matrix, and

$$\Lambda = \begin{bmatrix} \Lambda_r & 0 \\ 0 & 0 \end{bmatrix} \tag{37}$$

where

$$\Lambda_r = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \tag{38}$$

with $\lambda_i > 0 (i = 1, 2, \ldots r \leq m - l)$.

Let Q and Q^{-1} be partitioned as

$$Q = [Q_r \;\; Q_{m-r}] \quad Q^{-1} = \begin{bmatrix} Q_r^{-1} \\ Q_{m-r}^{-1} \end{bmatrix} \tag{39}$$

where Q_r, and Q_{m-r} are the first r and last $m - r$ columns of Q, respectively; in addition, Q_r^{-1}, and Q_{m-r}^{-1} are the first r and last $m - r$ rows of Q^{-1}, respectively. Then

$$\mathrm{I}_m = QQ^{-1} = Q_r Q_r^{-1} + Q_{m-r} Q_{m-r}^{-1} \tag{40}$$

and

$$PC = Q_r \Lambda_r Q_r^{-1}. \tag{41}$$

Now we have

$$\begin{aligned} \mathbf{w}(s) &= e^{-sPC}\mathbf{w}(0) = Qe^{-s\Lambda}Q^{-1}\mathbf{w}(0) \\ &= [Q_r \;\; Q_{m-r}] \begin{bmatrix} e^{-s\Lambda_r} & 0 \\ 0 & \mathrm{I}_{m-r} \end{bmatrix} \begin{bmatrix} Q_r^{-1} \\ Q_{m-r}^{-1} \end{bmatrix} \mathbf{w}(0) \\ &= \left[Q_r e^{-s\Lambda_r} Q_r^{-1} + Q_{m-r} Q_{m-r}^{-1} \right] \mathbf{w}(0) \end{aligned} \tag{42}$$

with

$$e^{-s\Lambda_r} = \begin{bmatrix} e^{-s\lambda_1} & & \\ & \ddots & \\ & & e^{-s\lambda_r} \end{bmatrix}. \tag{43}$$

Since all λ_i $(i = 1, 2, \ldots, r)$ are positive, the first term on the righthand side of Eq. 42 vanishes when $s \to \infty$, i.e.,

$$\mathbf{w}_\infty = \lim_{s\to\infty} \mathbf{w}(s) = Q_{m-r} Q_{m-r}^{-1} \mathbf{w}(0) \tag{44}$$

To establish $\mathbf{w}_\infty$ we need to determine Q_{m-r} and Q_{m-r}^{-1}.

Using Eqs. 40 and 41 we also have

$$\begin{aligned} \mathbf{w}_\infty &= Q_{m-r} Q_{m-r}^{-1} \mathbf{w}(0) = [\mathrm{I}_m - Q_r Q_r^{-1}]\mathbf{w}(0) \\ &= [\mathrm{I}_m - Q_r \Lambda_r \Lambda_r^{-1} Q_r^{-1}]\mathbf{w}(0) \\ &= [\mathrm{I}_m - Q_r \Lambda_r Q_r^{-1} Q_r \Lambda_r^{-1} Q_r^{-1}]\mathbf{w}(0) \\ &= \left[\mathrm{I}_m - PC(PC)^-\right]\mathbf{w}(0). \end{aligned} \tag{45}$$

Here we used the relations

$$\Lambda_r \Lambda_r^{-1} = \mathrm{I}_r, \tag{46}$$

$$Q_r^{-1} Q_r = \mathrm{I}_r, \tag{47}$$

$$\Lambda_r^{-1} = \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_r \end{bmatrix}, \tag{48}$$

and $(PC)^-$ denotes the generalized inverse of PC satisfying Penrose conditions (1) and (2). Considering that

$$\begin{aligned} Q_r Q_r^{-1} &= Q_r \Lambda_r^{-1} \Lambda_r Q_r^{-1} \\ &= Q_r \Lambda_r^{-1} Q_r^{-1} Q_r \Lambda_r Q_r^{-1} \\ &= (PC)^- PC, \end{aligned} \tag{49}$$

we have

$$PC(PC)^- = (PC)^- PC. \tag{50}$$

Thus, the other way to determine $\mathbf{w}_\infty$ is to find the matrix $(PC)^-$ that commutes with PC.

It can be proved that $\mathcal{K}_\infty$ is the minimum, i.e.,

$$\mathcal{K}(s) \geq \mathcal{K}_\infty = \lim_{s\to\infty} \mathcal{K}(s), \forall s \tag{51}$$

$$\begin{aligned}\mathcal{K}(s) &= \frac{1}{2}\mathbf{w}(s)^T C\mathbf{w}(s) = \frac{1}{2}\mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1} + Q_{m-r}Q_{m-r}^{-1}\right]^T \cdot \\ &\quad \cdot C\left[Q_r e^{-s\Lambda_r} Q_r^{-1} + Q_{m-r}Q_{m-r}^{-1}\right]\mathbf{w}(0) \\ &= \frac{1}{2}\mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]\mathbf{w}(0) \\ &\quad + \frac{1}{2}\mathbf{w}^T(0)\left[Q_{m-r}Q_{m-r}^{-1}\right]^T C\left[Q_{m-r}Q_{m-r}^{-1}\right]\mathbf{w}(0) \\ &\quad + \mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_{m-r}Q_{m-r}^{-1}\right]\mathbf{w}(0). \end{aligned} \tag{52}$$

The last term in Eq. 52 is zero (see Appendix A). Then we have

$$\begin{aligned}\mathcal{K}(s) &= \frac{1}{2}\mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]\mathbf{w}(0) + \frac{1}{2}\mathbf{w}_\infty^T C\mathbf{w}_\infty \\ &= \frac{1}{2}\mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]\mathbf{w}(0) + \mathcal{K}_\infty. \end{aligned} \tag{53}$$

Since the first term on the righthand side of Eq. 53 is nonnegative, we obtain Eq. 51, i.e., $\mathbf{w}_\infty$ is the solution of $\mathbf{w}$ with the smallest value of the cost function $\mathcal{K}$.

To determine $\mathbf{w}_\infty$ we can either calculate $(PC)^-$ or $Q_{m-r}Q_{m-r}^{-1}$. There is no general method to find a matrix $(PC)^-$ that commutes with PC. Direct determination of the eigenvalues and eigenvectors of PC may introduce a large error because PC is not symmetric. Therefore, we choose singular value decomposition to determine Q_{m-r} and Q_{m-r}^{-1}. Note that Q and Q^{-1} are the right and left eigenvector matrices of PC.

Using singular value decomposition [28], we have

$$PC = U\begin{bmatrix} A_r & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{54}$$

where U and V are orthonormal matrices with dimension m and

$$A_r = \begin{bmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_r \end{bmatrix} \tag{55}$$

with all $\alpha_i (i = 1, 2, \ldots, r) > 0$. Similarly, U and V can be partitioned as

$$U = [U_r \; U_{m-r}], \quad V = [V_r \; V_{m-r}]. \tag{56}$$

It is easy to see that the columns of V_{m-r} and U_{m-r}^T are the right and left eigenvectors of PC with the associated eigenvalue 0. However, these eigenvectors are not unique

because any linear combinations of the columns of V_{m-r} or U_{m-r}^T are still right and left eigenvectors of PC with respect to the eigenvalue 0. Thus, it is possible that

$$U_{m-r}^T V_{m-r} \neq \mathrm{I}_{m-r}. \tag{57}$$

We can set

$$Q_{m-r} = V_{m-r}, \tag{58}$$

but cannot simply set

$$Q_{m-r}^{-1} = U_{m-r}^T. \tag{59}$$

Suppose

$$U_{m-r}^T V_{m-r} = R \tag{60}$$

where R is an $(m-r)$-dimensional nonsingular square matrix. Then we see that

$$Q_{m-r}^{-1} = R^{-1} U_{m-r}^T \tag{61}$$

because

$$R^{-1} U_{m-r}^T V_{m-r} = R^{-1} R = \mathrm{I}_{m-r} \tag{62}$$

and

$$Q_{m-r} Q_{m-r}^{-1} = V_{m-r} R^{-1} U_{m-r}^T. \tag{63}$$

This method is stable because U and V resulting from the singular value decomposition are obtained from the symmetric matrix $(PC)^T PC$. When we use the solution of traditional least-squares regression for $\mathbf{w}$ as $\mathbf{w}(0)$, i.e., set $\mathbf{w}(0) = \mathbf{\Phi}^+ \mathbf{y}$, the final expression for $\mathbf{w}_\infty$ is

$$\mathbf{w}_\infty = V_{m-r} (U_{m-r}^T V_{m-r})^{-1} U_{m-r}^T \mathbf{\Phi}^+ \mathbf{y}. \tag{64}$$

Equation 64 is the key practical formula for the unknown parameter vector $\mathbf{w}$ obtained by D-MORPH regression. We can prove that the solution in Eq. 64 is unique for all $\mathbf{w}(0) \in \mathcal{M}$ (see Appendix B). As $\mathbf{\Phi}^+ \mathbf{y}$ is the solution of least-square regression, the new solution $\mathbf{w}_\infty$ given by D-MORPH regression is simply a linear combination of the elements of $\mathbf{w}$ obtained by least-squares regression. It can be readily proved that $\mathbf{w}_\infty$ preserves $\mathbf{y}$, i.e., satisfies $\mathbf{\Phi} \mathbf{w}_\infty = \mathbf{y}$ (see Appendix C).

For illustration, consider a simple function

$$y = \sin x \tag{65}$$

with data sets $(x^{(p)}, y^{(p)})$ (for $p \leq 4$, or $p \leq 5$). The regression model is set to be

Table 1 The initial and final values of the cost functions

Four data points				Five data points			
1st cost function		2nd cost function		1st cost function		2nd cost function	
$\mathcal{K}(0)$	$\mathcal{K}_\infty$	$\mathcal{K}(0)$	$\mathcal{K}_\infty$	$\mathcal{K}(0)$	$\mathcal{K}_\infty$	$\mathcal{K}(0)$	$\mathcal{K}_\infty$
8.5628	0.0000	27.9791	0.0001	7.7690	0.0347	21.7285	0.0000

Table 2 The initial and final values of the coefficients **w**

w_i	Four data points			Five data points		
		$\mathbf{w}_\infty$			$\mathbf{w}_\infty$	
	$\mathbf{w}(0)$	1st cost function	2nd cost function	$\mathbf{w}(0)$	1st cost function	2nd cost function
w_1	0.1006	0.0312	0.0360	0.0454	0.0195	0.0204
w_2	2.5597	2.9398	2.8951	2.2421	2.7538	2.7312
w_3	−1.2044	−0.3044	−0.2730	−0.3959	−0.1683	−0.0650
w_4	−1.0168	−3.3590	−2.9011	−0.3521	−2.8823	−2.7693
w_5	0.1269	0.0000	−0.5759	0.0974	0.0732	−0.6699
w_6	−0.9344	0.0000	−0.5160	−0.7960	0.0061	0.0886
w_7	0.4421	0.0000	0.5807	0.1680	0.0593	1.2811
w_8	−0.7404	0.0000	−0.3110	−0.8219	0.0057	−0.3526
w_9	0.5056	0.0000	0.2038	0.1645	0.0438	−0.5393
w_{10}	−0.6126	0.0000	0.6228	−0.7644	0.0045	0.1804

$$y = f(x) = \sum_{i=1}^{10} w_i x^{i-1}. \tag{66}$$

The D-MORPH method was used to search for the best solution. The first cost function is the weighted norm with

$$b_i = \begin{cases} 0, & \text{when } i \leq 4 \\ i, & \text{when } i > 4 \end{cases}$$

This setting keeps the coefficients for polynomials with degree less than 4 out of consideration for shrinkage. The second cost function is the weighted second order derivatives with the same b_i. The solution $\mathbf{w} = \mathbf{\Phi}^+\mathbf{y}$ obtained by least-square regression is used as $\mathbf{w}(0)$. The results are given in Tables 1, 2 and Fig. 1. Tables 1 and 2 show that the final values of the two cost functions are all smaller than their initial values, which implies that the weighted norm of **w** and the curvature of the fitted polynomial given by D-MORPH regression are smaller than those obtained by least-squares regression. When the weighted norm is used as the cost function, with four points of data ($p \leq 4$) the coefficients $w_i (i > 4)$ vanish. This corresponds to regression with only constant, linear, quadratic and cubic polynomials. For five points of data ($p \leq 5$),

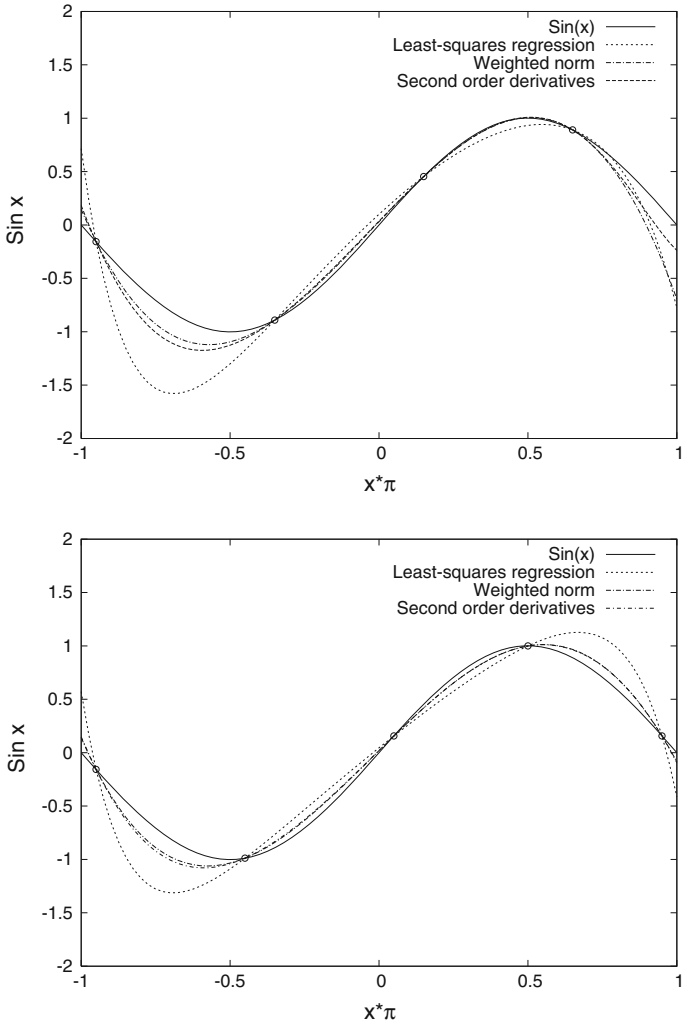

Fig. 1 Application of the D-MORPH regression method for a polynomial approximation of $y = \sin x$ obtained from four and five data points. The points on the curves are the data points

the coefficients $w_i (i > 4)$ do not vanish, but did shrink. Therefore, the contribution of high degree polynomials is either removed or reduced, and the variance of the model becomes smaller. In Fig. 1 the resultant two models obtained by D-MORPH regression are all better (i.e., they have a smaller prediction errors) than that obtained by direct least-squares regression.

3 Application to ignition of an H_2/air combustion model

An important aspect of combustion is ignition, which depends on several factors including the initial temperature, pressure and the equivalence ratio of the fuel and

oxygen. Experimental measurement of these variables is often time-consuming and costly, and typically only a limited amount of measured data can be obtained. Thus, construction of an accurate predictive model from limited data is desirable.

An H_2/air combustion model with 8 species (H_2, O_2, H_2O, H, O, OH, HO_2, H_2O_2) and 19 reactions [5] is used for testing the D-MORPH regression method. The initial temperature ($1,000 < T_0 < 1,500$ K), logarithmic value of pressure ($0.1 < P < 1$ atm), and logarithmic value of H_2/O_2 equivalence ratio ($0.1 < \phi < 10.0$) are chosen as three inputs denoted by (x_1, x_2, x_3), and the logarithmic value of homogeneous ignition delay (defined as the time lapse t_{ig} needed to attain an increase of 400 K from the initial temperature) is the output. One hundred random data points of $\mathbf{x}$ were sampled with a uniform distribution for T_0, $\log P$ and $\log \phi$ within the above ranges, and the corresponding logarithmic values of ignition delay, $\log t_{\mathrm{ig}}$, were calculated from the model. Gaussian white noise was added to the resultant $\log t_{\mathrm{ig}}$ to generate another set of data for the simulation of laboratory data. Both data sets (i.e., with and without noise) will be treated separately.

A 3rd order orthonormal polynomial expansion was used as the approximate model:

$$y(\mathbf{x}) = f_0 + \sum_{i=1}^{3}\sum_{r=1}^{3} \alpha_r^i \varphi_r^i(x_i) + \sum_{1\le i<j\le 3}\sum_{p=1}^{3}\sum_{q=1}^{3} \beta_{pq}^{ij} \varphi_p^i(x_i)\varphi_q^j(x_j) + \sum_{1\le i<j<k\le 3}\sum_{p=1}^{3}\sum_{q=1}^{3}\sum_{r=1}^{3} \gamma_{pqr}^{ijk} \varphi_p^i(x_i)\varphi_q^j(x_j)\varphi_r^k(x_k) \tag{67}$$

where the constant term f_0 is determined by the average value of the N measured output values

$$f_0 = \frac{1}{N}\sum_{j=1}^{N} y^{(j)} \tag{68}$$

and α_r^i, β_{pq}^{ij} and γ_{pqr}^{ijk} are constant coefficients to be determined; $\{\varphi\}$ are orthonormal polynomials

$$\varphi_1^i(x_i) = a_1 x_i + a_0, \tag{69}$$

$$\varphi_2^i(x_i) = b_2 x_i^2 + b_1 x_i + b_0, \tag{70}$$

$$\varphi_3^i(x_i) = c_3 x_i^3 + c_2 x_i^2 + c_1 x_i + c_0, \tag{71}$$

and fulfill the weighted orthonormality properties:

$$\int_0^1 w_i(x_i)\varphi_r^i(x_i)\mathrm{d}x_i \approx \frac{1}{N}\sum_{j=1}^{N} \varphi_r^i(x_i^{(j)}) \approx 0, \quad \text{for all } r, i \tag{72}$$

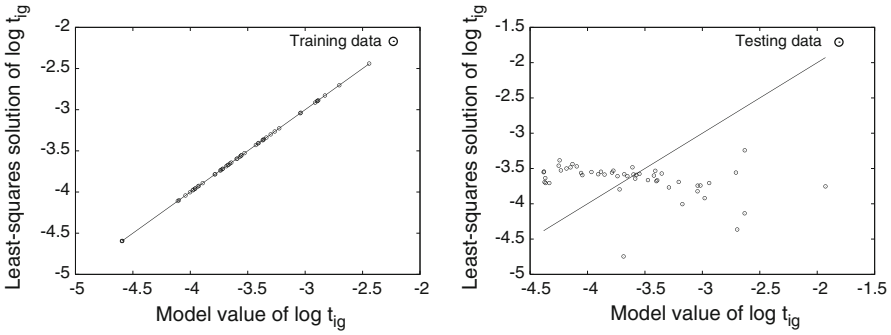

Fig. 2 The truth plot with least-squares regression for the training and testing data of the H_2/air combustion model without noise. Both the training and testing data sets contain 50 points

$$\int_0^1 w_i(x_i)[\varphi_r^i(x_i)]^2 \mathrm{d}x_i \approx \frac{1}{N}\sum_{j=1}^{N}[\varphi_r^i(x_i^{(j)})]^2 \approx 1, \quad \text{for all } r, i \tag{73}$$

$$\int_0^1 w_i(x_i)\varphi_p^i(x_i)\varphi_q^i(x_i)\mathrm{d}x_i \approx \frac{1}{N}\sum_{j=1}^{N}\varphi_p^i(x_i^{(j)})\varphi_q^i(x_i^{(j)}) \approx 0, \quad p \neq q \tag{74}$$

for a given set of data, where $w_i(x_i)$ is the probability density function for x_i, i.e., parameters $\{a_i, b_i, c_i\}$ are determined from solving Eqs. 72–74 [29–31].

3.1 Data set without noise

For the data set without noise, the values of $\log t_{\mathrm{ig}}$ are accurate and used as preserved features in D-MORPH regression, i.e., they will be fitted exactly. Using the 3rd order orthonormal polynomial expansion, there are 63 unknown parameters α_r^i, β_{pq}^{ij} and γ_{pqr}^{ijk} in Eq. 67. Fifty data points (i.e., the training data) were used to determine the parameters by least-squares regression. Another 50 points were used for testing the resultant model. The truth plot for the training and testing data by least-squares regression is given in Fig. 2. As the number of data (50) is less than the number of parameters (63), one can always find a set of α_r^i, β_{pq}^{ij} and γ_{pqr}^{ijk} to exactly fit the data. Thus, Fig. 2 shows that the fitting is exact for the training data, but the prediction quality for the testing data is meaningless, as expected.

Ridge regression given in Eq. 10 was used with the same training and testing data. Cross validation was employed to determine λ. The basic principle of cross validation is to leave the data points out one at a time and to choose the value of λ for which the missing data points are best predicted by the remainder of the data [17–19]. The resultant relation between the RSS of cross validation and parameter λ is given in Fig. 3. The minimum of the RSS is located at $\lambda = 0.00013$. The truth plot for the training and testing data obtained by ridge regression with $\lambda = 0.00013$ is given in Fig. 4.

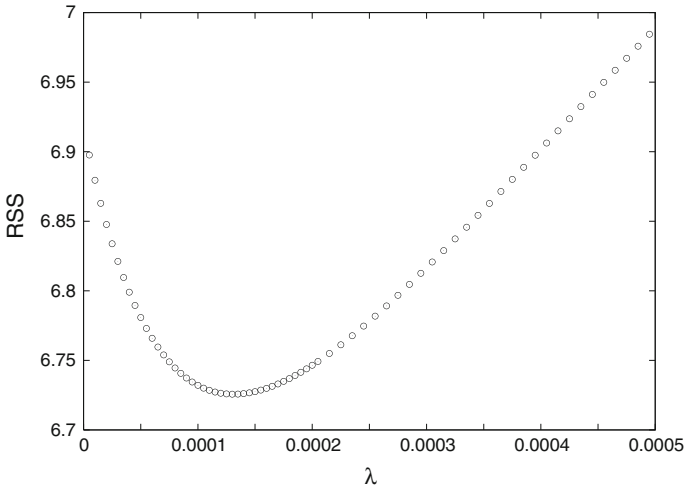

Fig. 3 The relationship between the RSS with cross validation and the parameter λ for ridge regression in the treatment of data without noise

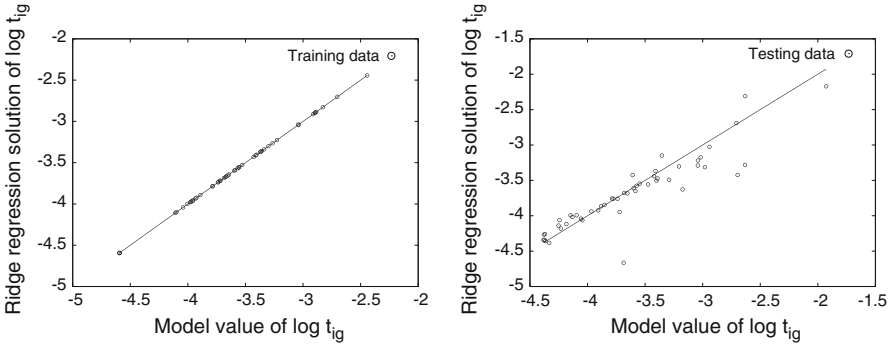

Fig. 4 The truth plot of the ridge regression method for the training and testing data of the H_2/air combustion model without noise. Both the training and testing data sets consist of 50 points

Since the λ is very small, its influence on fitting is negligible. Therefore, the fitting of ridge regression is accurate, and the prediction for the testing data is improved over Fig. 2. The average and maximum relative errors for the testing data are 4.42 and 26.93%, respetively. However, the computational effort is significant.

The D-MORPH regression method was then used with the same training and testing data. The weighted norm in Eq. 27 was used as the cost function. The weights for $\varphi_r^i(x_i)$, $\varphi_p^i(x_i)\varphi_q^j(x_j)$ and $\varphi_p^i(x_i)\varphi_q^j(x_j)\varphi_r^k(x_k)$ are set to be r, $p+q$ and $p+q+r$, respectively, i.e., the weight is equal to the sum of the degrees of the polynomials for each term. Therefore, the b_i's take the values between 1 and 9, and the higher degree polynomials have larger weights which makes them shrink faster. The rank of the projector P and the matrix PB cannot be larger than $m - N = 63 - 50 = 13$, and we found that both have rank 13. As B is diagonal, its eigenvalues are just the b_i's.

Table 3 The 13 nonzero eigenvalues of PB

i	$\lambda_i(B)$	$\lambda_i(PB)$	$\lambda_{50+i}(B)$	i	$\lambda_i(B)$	$\lambda_i(PB)$	$\lambda_{50+i}(B)$
1	9.0	6.34	3.0	8	7.0	4.41	2.0
2	8.0	6.00	3.0	9	7.0	4.20	2.0
3	8.0	5.59	3.0	10	7.0	3.94	2.0
4	8.0	5.34	3.0	11	6.0	3.60	1.0
5	7.0	5.10	2.0	12	6.0	3.42	1.0
6	7.0	4.82	2.0	13	6.0	3.08	1.0
7	7.0	4.46	2.0				

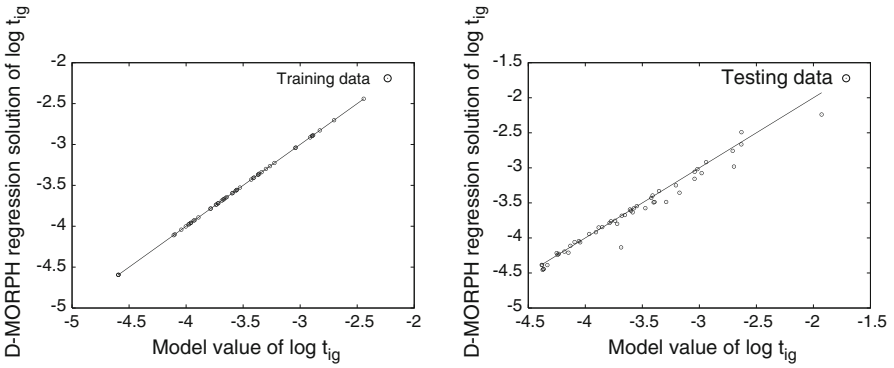

Fig. 5 The truth plot of D-MORPH regression for the training and testing data for the H_2/air combustion ignition model without noise. Both the training and testing data sets contain 50 points

Table 3 shows that the 13 nonzero eigenvalues of PB are all positive and satisfy the Poincaré separation theorem for eigenvalues.

The truth plot for the training and testing data obtained by the D-MORPH regression method is given in Fig. 5.

The exact fitting of the training data is preserved, and the prediction accuracy for the testing data is significantly improved over ridge regression. The average and maximum relative errors for the testing data are 1.95 and 16.13%, respetively. In this example, ridge regression shrinks all w_i's equally. In contrast, D-MORPH regression shrinks the w_i's for high order polynomials more to make the resultant model even smoother and has a better prediction accuracy. Even if ridge regression employs $\frac{1}{2}\mathbf{w}^T B\mathbf{w}$ as the regularization term, the reduction of fitting accuracy cannot be avoided. Moreover, the algebraic equation system, Eq. 5, needs to be solved only once in D-MORPH regression, and the computational effort is much smaller than ridge regression.

3.2 Data set with noise

The data set of $\log t_{\mathrm{ig}}$ with Gaussian white noise ε has average absolute and relative errors of 0.0627 and 1.75%, respectively. The signal to noise ratio is $\mathrm{Var}(\log t_{\mathrm{ig}})/\mathrm{Var}(\varepsilon)$

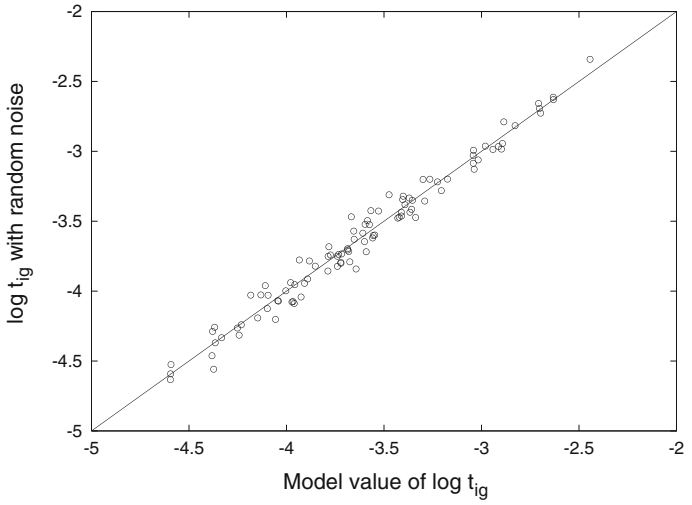

Fig. 6 The comparison between the two data sets for $\log t_{\mathrm{ig}}$ with and without noise

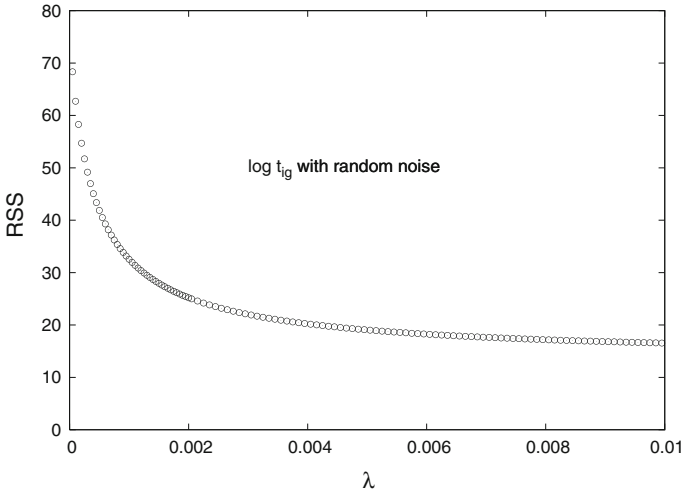

Fig. 7 The relationship between RSS with cross validation and the parameter λ for ridge regression in the treatment of the data with noise

≈ 43. Comparison between the two data sets for $\log t_{\mathrm{ig}}$ with and without noise is given in Fig. 6.

The 3rd order polynomial expansion, Eq. 67, is used for fitting in ridge regression. Both training and testing data sets still contain 50 data points, and cross validation is used to determine λ. The resultant relation between the RSS of cross validation and the parameter λ is given in Fig. 7. There is no minimum point, which is distinct from what was found in Fig. 3 obtained using the data without noise. In this case λ is chosen to be 0.004 where the RSS decrease varies slowly as λ increases. The truth plot for

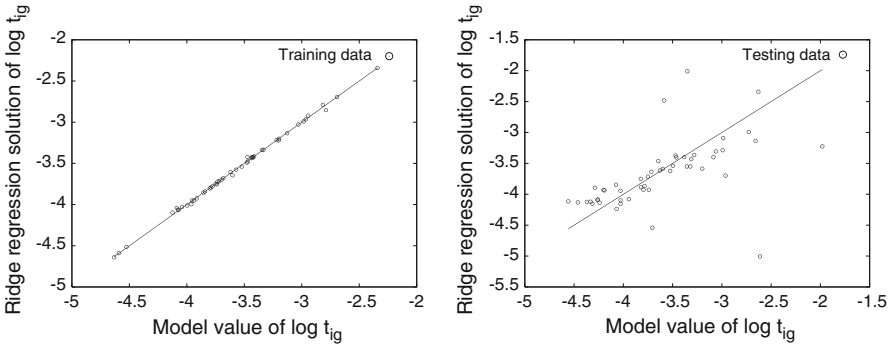

Fig. 8 The truth plot for the ridge regression method with the training and testing data for the H_2/air combustion model with noise. Both the training and testing data sets consist of 50 points

the training and testing data obtained by ridge regression with $\lambda = 0.004$ is given in Fig. 8. Since $\lambda = 0.004$ is small, the fitting of ridge regression is still good, but worse than that for the data without noise. However, the prediction accuracy for the testing data is significantly worse. The average and maximum relative errors for the testing data are 9.79 and 91.58%, respetively.

When $\log t_{\rm ig}$ has error, it is improper to choose the data $\mathbf{y}$ as the preserved features in D-MORPH regression for exact fitting. Doing so would diminish the prediction accuracy of the resultant model. Under these conditions a general approach is to filter the data $\mathbf{y}$ to produce a meta-data set $\hat{\mathbf{y}}$ used as the preserved features in D-MORPH regression. There are many ways to accomplish this filtering. The data set with noise may be fitted by a simpler model composed of a proper subset of basis polynomial functions in Eq. 67 such that the number of unknown parameters is less than the number of data thereby acting as a filter. For example, with 50 training data the following two models with 44 and 36 unknown parameters

$$\hat{y}_1(\mathbf{x}) = f_0 + \sum_{i=1}^{3}\sum_{r=1}^{3}\alpha_r^i\varphi_r^i(x_i) + \sum_{1\le i<j\le 3}\sum_{p=1}^{3}\sum_{q=1}^{3}\beta_{pq}^{ij}\varphi_p^i(x_i)\varphi_q^j(x_j) + \sum_{1\le i<j<k\le 3}\sum_{p=1}^{2}\sum_{q=1}^{2}\sum_{r=1}^{2}\gamma_{pqr}^{ijk}\varphi_p^i(x_i)\varphi_q^j(x_j)\varphi_r^k(x_k), \tag{75}$$

$$\hat{y}_2(\mathbf{x}) = f_0 + \sum_{i=1}^{3}\sum_{r=1}^{3}\alpha_r^i\varphi_r^i(x_i) + \sum_{1\le i<j\le 3}\sum_{p=1}^{3}\sum_{q=1}^{3}\beta_{pq}^{ij}\varphi_p^i(x_i)\varphi_q^j(x_j) \tag{76}$$

may be used to serve as data noise filtering meta-models for least-squares regression. However, such arbitrary choices of a subset of basis polynomial functions in Eq. 67 may not be best for this purpose. Thus, a general and systematic method will also be employed using a statistical F-test to identify the significant polynomial basis functions in Eq. 67 for a given set of data to compose a simple model acting as a filter. As the simple model is determined from a statistical test, the deviation of the model

prediction value $\hat{y}$ from y is just the random error in y [9]. The predictions based on this filtered meta-data $\hat{\mathbf{y}}$ should be closer to the true value of $\mathbf{y}$ and can be used as preserved features in D-MORPH regression. We will also retain the models in Eqs. 75 and 76 for comparison.

Let g_l denote any polynomial basis function in Eq. 67. Suppose that RSS_1 represents the residual sum-of-squares for the least-squares regression of a large model $f_0 + \sum_{l=1}^{r} g_l$ with p_1 parameters, and RSS_0 is the same quantity for a small model $f_0 + \sum_{l=1,l\neq k}^{r} g_l$ nested in the large model, but with p_0 parameters. The F statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1)} \tag{77}$$

has a F distribution with $(p_1 - p_0)$ and $(N - p_1)$ degrees of freedom. If the observed F given by Eq. 77 is larger than the tabulated value of the F distribution with $(p_1 - p_0)$ and $(N - p_1)$ degrees of freedom at the 99% confidence level (or other desired confidence level), then g_k is significant and should be included in the approximation. Otherwise, g_k can be excluded. Forward or backward stepwise selection may be used to search for significant polynomial basis functions. For the 50 points of training data with noise, the resultant simpler model was found to be

$$\begin{aligned}\hat{y}_3(\mathbf{x}) = f_0 + \sum_{i=1}^{3} \alpha_1^i \varphi_1^i(x_i) + \alpha_2^3 \varphi_2^3(x_3) \\ + \gamma_{111}^{123} \varphi_1^1(x_1)\varphi_1^2(x_2)\varphi_1^3(x_3).\end{aligned} \tag{78}$$

The resultant $\hat{y}_3(\mathbf{x}^{(j)})$ $(j = 1, 2, \ldots, 50)$ are used to replace $y(\mathbf{x}^{(j)})$ (i.e., the training data of $\log t_{\mathrm{ig}}$ with noise) as preserved features in D-MORPH regression with Eq. 67. Since the number (63) of unknown parameters in Eq. 67 is larger than the number of data, there is an infinite number of solutions for $\mathbf{w}$ that will exactly fit the $\hat{y}_3(\mathbf{x}^{(j)})$. Note that the solution $\mathbf{w}$ of Eq. 78 obtained by least-squares regression belongs to these solutions if the extra elements of $\mathbf{w}$ introduced in Eq. 67 are considered to be zero. However, the cost function $\mathcal{K}$ (the weighted norm for $\mathbf{w}$) has the smallest value for the solution with D-MORPH regression, i.e., smaller than the cost function value for the solution of Eq. 78 given by least-squares regression. Therefore, after $y(\mathbf{x}^{(j)})$ is replaced by $\hat{y}_3(\mathbf{x}^{(j)})$, D-MORPH regression with Eq. 67 will have the same fitting accuracy for the training data and better prediction accuracy for the testing data compared to least-squares regression for Eq. 78. This is also true for Eqs. 75 and 76 because the polynomial basis functions used in Eqs. 75 and 76 are subsets of the polynomial basis functions used in Eq. 67. Table 4 compares the values of cost function, average absolute errors, and average relative errors for the testing data. The arbitrary choices of filtering meta-models in Eqs. 75 and 76 are also retained for comparison with the statistically based optimal method Eq. 78.

The results in Table 4 show that: (1) for least-squares regression Eq. 78 has the best prediction accuracy, which demonstrates that Eq. 78 is an appropriate filter; (2) D-MORPH regression has equal or lower values of the cost function, average absolute and average relative errors than least-squares regression no matter which simple

Table 4 The comparison of the values of cost function, average absolute and relative errors for least-squares regression (LS) and D-MORPH (DM) regression with testing data

Replacement of y	$\mathcal{K}$		Ave. abs. err.		Ave. rel. err. (%)	
	LS	DM	LS	DM	LS	DM
$\hat{y}_1$	1.36	0.53	0.21	0.14	6.38	4.11
$\hat{y}_2$	0.25	0.19	0.17	0.15	5.21	4.52
$\hat{y}_3$	0.14	0.14	0.11	0.11	3.28	3.47

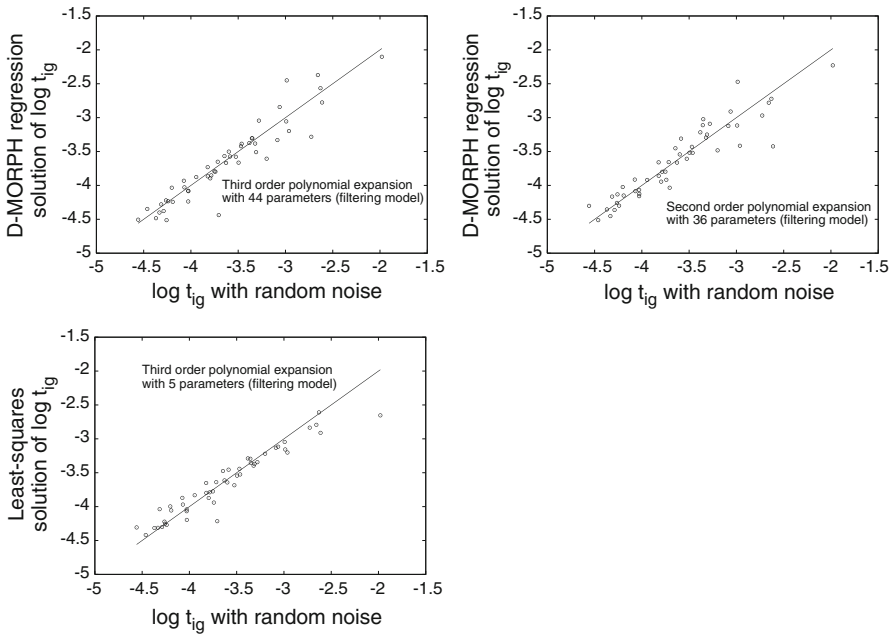

Fig. 9 The truth plot of D-MORPH regression for the testing data with noise of the H_2/air combustion ignition model based on use of $\hat{y}_1$, $\hat{y}_1$ and $\hat{y}_3$

model is used as a filter. For Eq. 78 all $\varphi_r^i(x_i)$'s are linear in x_i (except for the quadratic function $\varphi_2^3(x_3)$), which implies that there is little possibility to further smooth by any regularization. This is not common for other systems and data sets. Thus, D-MORPH and least-squares regressions have the same value (the average relative error given by D-MORPH regression is a little larger, but the difference is not significant for only 50 points). Therefore, D-MORPH regression always has the best prediction accuracy while the fitting accuracy is preserved. The truth plot for the testing data obtained by D-MORPH regression based on $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$ is given in Fig. 9 which is better than Fig. 8 obtained from ridge regression.

4 Conclusion

This paper utilizes D-MORPH for linear basis function regression with more parameters than observation data. D-MORPH regression performs an exploration of the parameters $\mathbf{w}$ in the manifold $\mathcal{M}$ (composed of all the solutions $\mathbf{w}$ of the consistent linear algebraic equation system) with all training data being exactly fit and an additional cost function continuously reduced. The additional cost function $\mathcal{K}$ is used to reduce the model variance and consequently improve the prediction accuracy. The paper considered a cost function $\mathcal{K} = \frac{1}{2}\mathbf{w}^T C\mathbf{w}$ with the matrix C being positive or semi-positive definite. This cost is a common form, and under this condition the differential equation for the exploration path is linear and a closed form unique solution for $\mathbf{w}$ can be obtained corresponding to the global minimum of cost function. The resultant D-MORPH regression parameters are linear combinations of the parameters given by least-squares regression. When the output data $\mathbf{y}$ has noise, the data $\mathbf{y}$ is filtered and replaced by the meta-data $\hat{\mathbf{y}}$ from a simpler model constructed using a statistical F-test with less unknown parameters than the number of data. Then the treatment follows as for the data without noise. The illustrative examples demonstrate that D-MORPH regression is better than least-squares and ridge regression. D-MORPH regression can be extended to costs $\mathcal{K}(\mathbf{w})$ of more complex form than quadratic, and then a numerical trajectory over $\mathbf{w}(s)$ could be performed.

Acknowledgments The authors thank Vincent Beltrani for helpful discussions with programming guidance. Support for this work has been provided partially by the USEPA through the Center for Exposure and Risk Modeling (CERM—EPAR827033) and the Environmental Bioinformatics and Computational Toxicology Center (ebCTC—GAD R 832721-010) and the Department of Energy.

Appendices

Appendix A: $\mathbf{w}_\infty$ has the smallest value of $\mathcal{K}$

To prove that $\mathbf{w}_\infty$ has the smallest value of $\mathcal{K}$ we need to show that

$$\mathbf{w}^T(0)\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_{m-r} Q_{m-r}^{-1}\right]\mathbf{w}(0) = 0. \tag{79}$$

It suffices to prove that $\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C\left[Q_{m-r} Q_{m-r}^{-1}\right]$ is a null matrix. First, using Eq. 41 we expand

$$\begin{aligned}\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right] &= Q_r\left[\mathrm{I}_r - s\Lambda_r + \frac{s^2}{2!}\Lambda_r^2 - \ldots\right]Q_r^{-1}\\ &= Q_r Q_r^{-1} - sQ_r\Lambda_r Q_r^{-1} + \frac{s^2}{2!}Q_r\Lambda_r^2 Q_r^{-1} - \cdots\\ &= Q_r\Lambda_r Q_r^{-1} Q_r\Lambda_r^{-1} Q_r^{-1} - sQ_r\Lambda_r Q_r^{-1}\\ &\quad + \frac{s^2}{2!}Q_r\Lambda_r Q_r^{-1} Q_r\Lambda_r Q_r^{-1} - \cdots\end{aligned}$$

$$= PC(PC)^{-} - sPC + \frac{s^2}{2!}(PC)^2 - \cdots . \tag{80}$$

Similarly

$$\begin{aligned} Q_{m-r}Q_{m-r}^{-1} &= \mathrm{I}_m - Q_r Q_r^{-1} = \mathrm{I}_m - Q_r \Lambda_r^{-1}\Lambda_r Q_r^{-1} \\ &= \mathrm{I}_m - Q_r \Lambda_r^{-1} Q_r^{-1} Q_r \Lambda_r Q_r^{-1} \\ &= \mathrm{I}_m - (PC)^{-} PC. \end{aligned} \tag{81}$$

Now we have

$$\begin{aligned} &\left[Q_r e^{-s\Lambda_r} Q_r^{-1}\right]^T C \left[Q_{m-r} Q_{m-r}^{-1}\right] \\ &= \left[PC(PC)^{-} - sPC + \frac{s^2}{2!}(PC)^2 - \cdots\right]^T C \left[\mathrm{I}_m - (PC)^{-} PC\right] \\ &= \left[((PC)^{-})^T CP - sCP + \frac{s^2}{2!}(CP)^2 - \cdots\right] C \left[\mathrm{I}_m - (PC)^{-} PC\right] \\ &= \left[((PC)^{-})^T CPC - sCPC + \frac{s^2}{2!} CPCPC - \cdots\right] \left[\mathrm{I}_m - (PC)^{-} PC\right] \\ &= \left[((PC)^{-})^T CPC - sCPC + \frac{s^2}{2!} CPCPC - \cdots\right] \\ &\quad - \left[((PC)^{-})^T CPC(PC)^{-} PC - sCPC(PC)^{-} PC \right. \\ &\quad \left. + \frac{s^2}{2!} CPCPC(PC)^{-} PC - \cdots\right] \\ &= \left[((PC)^{-})^T CPC - sCPC + \frac{s^2}{2!} CPCPC - \cdots\right] \\ &\quad - \left[((PC)^{-})^T CPC - sCPC + \frac{s^2}{2!} CPCPC - \cdots\right] = 0. \end{aligned} \tag{82}$$

Here we used the symmetric property of P and C, and

$$PC(PC)^{-} PC = PC. \tag{83}$$

Appendix B: The uniqueness of the solution given in Eq. 64

First we prove that the manifold $\mathcal{M} \subset \Re^m$ composed of the solutions of Eq. 5 is completely connected. Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{M}$, i.e., they satisfy

$$\mathbf{\Phi}\mathbf{w}_1 = \mathbf{y}, \qquad \mathbf{\Phi}\mathbf{w}_2 = \mathbf{y}. \tag{84}$$

Then $\lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2 \in \mathcal{M}$ with $0 \le \lambda \le 1$ because

$$\begin{aligned} \mathbf{\Phi}[\lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2] &= \lambda \mathbf{\Phi}\mathbf{w}_1 + (1-\lambda)\mathbf{\Phi}\mathbf{w}_2 \\ &= \lambda \mathbf{y} + (1-\lambda)\mathbf{y} = \mathbf{y} \end{aligned} \tag{85}$$

which implies that $\mathcal{M}$ is a convex set and any two points of $\mathcal{M}$ are connected.

Suppose two initial $\mathbf{w}(0)$'s

$$\mathbf{w}_1(0) = \mathbf{\Phi}^+\mathbf{y} + (\mathrm{I}_m - \mathbf{\Phi}^+\mathbf{\Phi})\mathbf{v}_1, \tag{86}$$

$$\mathbf{w}_2(0) = \mathbf{\Phi}^+\mathbf{y} + (\mathrm{I}_m - \mathbf{\Phi}^+\mathbf{\Phi})\mathbf{v}_2 \tag{87}$$

satisfying Eq. 5 are used. Since $\mathbf{v}_1$ and $\mathbf{v}_2$ are arbitrary vectors in $\Re^m$, the initial $\mathbf{w}_1(0)$ and $\mathbf{w}_2(0)$ are two arbitrary points in the manifold $\mathcal{M}$. Their corresponding solutions for $\mathbf{w}_\infty$ represented by Eq. 45 are

$$\mathbf{w}_{\infty 1} = \left[\mathrm{I}_m - PC(PC)^-\right]\mathbf{w}_1(0), \tag{88}$$

$$\mathbf{w}_{\infty 2} = \left[\mathrm{I}_m - PC(PC)^-\right]\mathbf{w}_2(0). \tag{89}$$

The difference of the two solutions is

$$\begin{aligned} \mathbf{w}_{\infty 1} - \mathbf{w}_{\infty 2} &= \left[\mathrm{I}_m - PC(PC)^-\right](\mathbf{w}_1(0) - \mathbf{w}_2(0)) \\ &= \left[\mathrm{I}_m - PC(PC)^-\right](\mathrm{I}_m - \mathbf{\Phi}^+\mathbf{\Phi})(\mathbf{v}_1 - \mathbf{v}_2) \\ &= \left[\mathrm{I}_m - PC(PC)^-\right]P(\mathbf{v}_1 - \mathbf{v}_2) \\ &= \left[P - PC(PC)^-P\right](\mathbf{v}_1 - \mathbf{v}_2) \\ &= \left[P - PC(PC)^-PCC^{-1}\right](\mathbf{v}_1 - \mathbf{v}_2) \\ &= \left[P - PCC^{-1}\right](\mathbf{v}_1 - \mathbf{v}_2) \\ &= [P - P](\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}. \end{aligned} \tag{90}$$

Here C is assumed to be nonsingular. The difference between the two solutions being zero implies that starting from an arbitrary point $\mathbf{w}(0) \in \mathcal{M}$ will result in the same $\mathbf{w}_\infty$. Since $\mathcal{M}$ is completely connected, the solution is unique. As $\mathcal{K}(\mathbf{w}_\infty) \le \mathcal{K}(\mathbf{w}(s))$, then $\mathcal{K}(\mathbf{w}_\infty)$ is the global minimum of $\mathcal{K}$ in the entire manifold $\mathcal{M}$.

Appendix C: $\mathbf{w}_\infty$ preserves $\mathbf{y}$

$\mathbf{w}_\infty$ preserves $\mathbf{y}$ if

$$\mathbf{\Phi}\mathbf{w}_\infty = \mathbf{y}. \tag{91}$$

Using Eq. 45

$$\begin{aligned}\mathbf{\Phi}\left[\mathrm{I}_m - PC(PC)^{-}\right]\mathbf{w}(0) &= \mathbf{\Phi}\left[\mathrm{I}_m - (\mathrm{I}_m - \mathbf{\Phi}^{+}\mathbf{\Phi})C(PC)^{-}\right]\mathbf{w}(0)\\ &= \left[\mathbf{\Phi} - (\mathbf{\Phi} - \mathbf{\Phi}\mathbf{\Phi}^{+}\mathbf{\Phi})C(PC)^{-}\right]\mathbf{w}(0)\\ &= \left[\mathbf{\Phi} - (\mathbf{\Phi} - \mathbf{\Phi})C(PC)^{-}\right]\mathbf{w}(0)\\ &= \mathbf{\Phi}\mathbf{w}(0) = \mathbf{y}. \end{aligned} \tag{92}$$

References

1. D.C. Peters, F.R. Korosec, T.M. Grist, W.F. Block, J.E. Holden, K.K. Vigen, C.A. Mistretta, Magn. Reson. Med. **43**, 91–101 (2000)
2. B.L. Daniel, Y.F. Yen, G.H. Glover et al., Radiology **209**, 499–509 (1998)
3. E.J. Candés, J. Romberg, in Proceedings of SPIE International Symposium on Electro. Imaging **1**, 76–86 (2005), San Jose
4. A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, J.S. Brosy, Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc. Nat. Acad. Sci. USA **101**(27), 10143–10148 (2004)
5. J. Li, Z.W. Zhao, A. Kazakov, F.L. Dryer, Int. J. Chem. Kinet. **36**, 566–575 (2004)
6. J. Kettenring, B. Lindsay, D. Siegmud (eds). Statistics: Challenges and opportunities for the twenty-first century. NSF report. Available at www.pnl.gov/scales/docs/nsf_report.pdf
7. E. Cades, T. Tao, Ann. Statist. **35**, 2313 (2007)
8. C.M. Carvalho, J.C. Chang, J.E. Lucas, J.R. Nevins, Q. Wang, M. West, Am. Stat. Assoc. **103**, 1438 (2008)
9. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, New York, 2001)
10. C.M. Bishop, *Pattern recognition and machine learning* (Springer, New York, 2007)
11. A.N. Tikhonov, Dokl. Akad. Nauk. SSSR **39**, 195–198 (1943)
12. A.N. Tikhonov, Soviet Math. Dokl. **4**, 1035–1038(1963). English translation of Dokl. Akad. Nauk. SSSR **151**, 501–504 (1963)
13. A.N. Tikhonov, V.A. *Arsenin, Solution of Ill-posed Problems*. Winston & Sons, Washington, (1977). ISBN 0-470-99124-0.
14. P.C. Hansen, Rank-deficient and Discrete ill-posed problems. (1998), SIAM.
15. A.E. Hoerl, Chem. Eng. Prog. **58**, 54–59 (1962)
16. A.E. Hoerl, R. Kennard, Technometrics **12**, 55–67 (1970)
17. G. Wahba, *Spline models for observational data* (SIAM, Philadelphia, 1990)
18. G. Wahba, Ann. Stat. **13**, 1378–1402 (1985)
19. G. Wahba, Y.D. Wang, C. Gu, R. Klein, B. Klein, Ann. Stat. **23**, 1865–1895 (1995)
20. http://en.wikipedia.org/wiki/Tikhonov_regularization, Categories: Linear algebra, Estimation theory Views.
21. A. Rothman, T.-S. Ho, H. Rabitz, Phys. Rev. A **72**, 023416 (2005)
22. A. Rothman, T.-S. Ho, H. Rabitz, J. Chem. Phys. **123**, 134104 (2005)
23. A. Rothman, T.-S. Ho, H. Rabitz, Phys. Rev. A **73**, 053401 (2006)
24. N. Danielson, V. Beltrani, J. Dominy, H. Rabitz, (manuscript in preparation)
25. C.R. Rao, S.K. Mitra, *Generalized inverse of matrix and its applications* (Willey, New York, 1971)
26. Matlab [7.0R14], 2004. MathWorks, Inc
27. R. Bellman, *Introduction to matrix analysis. 2nd edn* (McGraw-hill, New York, 1970), p. 118
28. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical recipes in FORTRAN—The art of science computing. 2nd edn* (Cambridge university press, New York, 1992), p. 51
29. O.F. Alis, H. Rabitz, J. Math. Chem. **25**, 197–233 (1999)
30. G. Li, S.W. Wang, H. Rabitz, J. Phys. Chem. A **106**, 8721–8733 (2002)
31. G. Li, J.S. Hu, S.W. Wang, P.G. Georgopoulos, J. Schoendorf, H. Rabitz, J. Phys. Chem. A **110**, 2474–2485 (2006)